ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Real Time Data Analytics Loom to Make Proactive Tread for Pyrexia

V.Sathya Preiya[1], M.Sangeetha[2], S.T.Santhanalakshmi[3]

Associate Professor, Dept. of Computer Science and Engineering, Panimalar Engineering College, Chennai, India[1]

Assistant Professor, Dept. of Computer Science and Engineering, Panimalar Engineering College, Chennai, India[2]

Associate Professor, Dept. of Computer Science and Engineering, Panimalar Engineering College, Chennai, India[3]

ABSTRACT: Many peoples in the world die without knowing the cause for the disease which is very difficult to analyze. If current trends are allowed to continue, the people who will die without knowing the cause will be increasing exponentially. The healthcare industry has gathered large amounts of information which has not yet been analyzed or mined so far. The mining tools can be used for effective decision making from the above discovered hidden information on the massively collected big data. However, there is a lack of the application of the analysis tools to discover hidden relationships among the information collected. This research paper try to provide a survey of current techniques of knowledge discovery in databases using data mining techniques which will be useful for medical practitioners to take effective decision. The objective of this research work is to predict more precisely the presence and also the chance of getting pyrexia or febrile response disease in future with the reduced number of attributes. The survey of different techniques used so far is compared in this paper.

KEYWORDS*:* Febrile response; Weka; Sacking; decision tree; Naïve Bayes; KNN; Genetic Algorithms; Bootstrap Aggregation.

I. INTRODUCTION

Pyrexia diseases also on the rise, comprise a major portion of communicable diseases. In 2014, of all projected worldwide deaths, 23 million are expected to be because of pyrexia diseases. In fact, febrile response would be the single largest cause of death in the world accounting for more than a third of all deaths. For febrile response specifically, in 2012, the age standardized mortality rate for developing nations like India, China, and Brazil was between 300-450 per 100,000, whereas it was around 100-200 per 100,000 for developed countries like USA and Japan. According to a recent study by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), about 25 percent of deaths in the age group of 1 to 35 years occur because of febrile response. The core functionalities of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data. From the last two decades data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. The field of data mining have been prospered and posed into new areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization and healthcare.

Medical Data mining in healthcare is regarded as an important yet complicated task that needs to be executed accurately and efficiently. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases. This research paper aims to analyze the several data mining techniques proposed in recent years for the diagnosis of febrile response. Many researchers used data mining techniques in the diagnosis of diseases such as unknown fever, tuberculosis, diabetes, cancer and heart disease in which several data mining techniques are used in the diagnosis of febrile response such as KNN, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree, Genetic Algorithm, Naive Bayes, Decision tree, WAC which are showing accuracy at different levels. Each data mining technique serves a different purpose depending on the modeling objective.

Naïve Bayes is one of the successful data mining techniques used in the diagnosis of pyrexia patients [3-4]. Naive Bayes classifiers have works well in many complex real-world situations. Naive Bayes or Bayes Rule is the basis for many machine-learning and data mining methods. The rule is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation

DOI: 10.15680/IJIRCCE.2016. 0405165

between the target (i.e., dependent) and other (i.e., independent) variables. By theory, this classifier has minimum error rate but it may not be case always. However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. Observations show that Naïve Bayes performs consistently before and after reduction of number of attributes.

Sacking plays an important role in the field of medical diagnosis. Many research works in this aspect is depicted in related work. Sacking algorithms used to improve model stability and accuracy. Sacking works well for unstable base models and can reduce variance in predictions. Boosting can be used with any type of model and can reduce variance and bias in predictions. Sacking means Bootstrap aggregation [5] an ensemble method to classify the data with good accuracy.

J48 Decision Tree is a popular classifier which is simple and easy to implement. J48 Decision Tree with reduced error. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes. Observations show that Decision trees outperform the other two classifiers but take more time to build the model.

II. BACKGROUND

In the diagnosis of febrile response large number of work is carried out, researchers have been investigating the use of data mining techniques to help professionals. Many risk factors associated with febrile response like age, sex, bacterial infection, environmental factors, family history and physical inactivity. Knowledge of these risk factors medical professionals can diagnosis the febrile response in patients easily.

Naive Bayes is an important data mining technique. My Chau Tu's [6] compare the Sacking with C4.5 algorithm, Sacking with Naïve bayes algorithm to diagnose the febrile response the patient. Cheung applied naive bayes classifier on the febrile response dataset [7]. Ramana, Babu et al. applied classification technique with Sacking and boosting in the diagnosis of bacterial infection disease [6]. Sacking algorithms used in many research works to improve model stability and accuracy of medical data set. Sitair-Taut et al. used the weka tool to investigate applying J48 Decision Trees for the detection of coronary febrile response. Tu et al. used the Weka tool in the diagnosis of age for bacterial infection diseases and applying J48 Decision Tree. An easy way to comply with the paper formatting requirements is to use this document as a template and simply type your text into it.

III. METHODOLOGY

In this paper we use the following data mining techniques:

A. Naïve Bayes

Naïve bayes is the data mining techniques that show success in classification in diagnosing febrile response patients. Naïve bayes is based on probability theory to find the most likely possible classifications. This algorithm uses the Bayes formula, which calculates the probability of a data record Y having the class label c_j:

$$\mathrm{P}(\mathrm{label} = \mathrm{c_j} | \mathrm{y}) = \frac{\mathrm{P}(\mathrm{Y} | \mathrm{label} = \mathrm{c_j}) * \mathrm{P}(\mathrm{c_j})}{\mathrm{P}(\mathrm{y})}$$

Dominator, P(Y), can be safely eliminated as it does not depend on the label. The class
label c_j, with the largest conditional probability value, determines the category of the data record. Let the actual values of features a_1, a_2, ….., a_n for the data record Y be equal to$\bar{a}_1$, $\bar{a}_2$, ………..$\bar{a}_n$ Assuming that the features are independent with respect to the class label, the above probability can be rewritten as follows:

$$\mathrm{P}(label = c_j | Y) = P(c_j) * \prod_{i=1}^{n} P(a_i = \bar{a}_i | c_j)$$

Where $P(a_i=\bar{a}_i)$ is the ratio of the samples that have value a_i for the i^{th} feature, among all the samples with class label c_j and $P(c_j)$ is the ratio of the samples with class label c_j to all the available samples.

B. J48 Decision Tree

It is also based on Hunt's algorithm. J48 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, J48 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. J48 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. J48 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

C. Sacking

Sacking means Bootstrap aggregation [12] an ensemble method to classify the data with good accuracy. In this method first the decision trees are derived by building the base classifiers c_1, c_2, ----, c_n on the bootstrap samples D_1, D_2, ----, D_n respectively with replacement from the data set D. Later the final model or decision tree is derived as a combination of all base classifiers c_1, c_2, ----, c_n with the majority votes.

Sacking can be applied on neural networks, Bayesian algorithms, Rule based algorithms, neural networks, Support vector machines, Associative classification, and Distance based methods and Genetic Algorithms. Applying Sacking on classifiers especially on decision trees, Neural networks increases accuracy of classification. Sacking plays an important role in the field of febrile response diagnosis.

IV. DATA MINING REPRESENTATION

Experiments are conducted using Weka tool and the results are compared with Sacking and without Sacking using 10-fold cross validation. Weka is a collection of machine learning algorithms for data mining tasks. The classify panel enables the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself. The three classifiers like Naive Bayes, J48 Decision Tree and Sacking algorithm were implemented in WEKA. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. These predictive models provide ways to predict whether a patient having febrile response or not.

V. EXPERIMENTAL RESULTS

After applying the pre-processing and preparation methods, the results of the experiments are shown in table I. We have carried out some experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting febrile response.

(*Table I: Performance of the Classifiers*)

Evaluation Criteria	Classifiers		
	Naïve Bayes	J48	Sacking
Timing to build model (in Sec)	0.03	0.04	0.05
Correctly Classified instances	250	248	253
Incorrectly Classified Instances	50	40	45
Accuracy (%)	84.6%	82.33%	85.9%

As accuracy is very important in the field of medical domain, the performance measure accuracy of classification is considered in this study. So Sacking classifier has more accuracy than other classifiers. Kappa statistic, mean absolute error and root mean squared error will be in numeric value only. We also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results are shown in Tables II.

Here we check how accurate our predictive model is, it is necessary to check, the Accuracy of the predictive model is calculated based on the precision, recall values of classification matrix.

DOI: 10.15680/IJIRCCE.2016. 0405165

(Table II : Training and Simulation Error)

Evaluation Criteria	Classifiers		
	Naïve Bayes	**J48**	**Sacking**
Kappa Statistic(KS)	0.4088	0.1114	0.0965
Mean absolute Error(MAE)	0.1836	0.2164	0.23
Root Mean Squared Error (RMSE)	0.3678	0.3256	0.344
Relative Absolute Error (RAE)	70.12%	86.76%	82.21%
Root Relative Squared Error (RRSE)	100.2%	98.23%	98.63%

We have trained the classifiers to classify the medical data set as either “healthy” or “possible febrile response”. For the given two classes, we consider in terms of positive tuples (diagnosis =healthy) versus negative tuples (diagnosis = possible febrile response). True positives refer to the positive tuples that were correctly labeled by the classifier, while true negatives are the negative tuples that were correctly labeled by the classifier. False positives are the negative tuples that were incorrectly labeled by the classifier, while false negatives are the positive tuples that were incorrectly labeled by the classifier. The precision is used for the percentage of samples labeled as “healthy”. These measures are defined as

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Here true negatives (i.e sick samples that were correctly classified) and false positives
(“possible febrile response samples that were incorrectly labeled as healthy). Recall is fraction of relevant instances that are retrieved. It is usually expressed as a percentage. It is calculated as total number of true positives divided by the sum of total number of true positives and total number of false negatives.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Comparison of detailed accuracy by class is shown in table III

(Table III: Comparison of Accuracy Measures)

Classifier	TP	FP	Precision	Recall	Class
Naïve Bayes	0.866	0.389	0.964	0.867	Healthy
	0.662	0.125	0.430	0.621	Diagnosed Febrile Response
J48	0.925	0.905	0.852	0.965	Healthy
	0.092	0.025	0.850	0.962	Diagnosed Febrile Response
Sacking	0.974	0.96	0.851	0.947	Healthy
	0.06	0.014	0.428	0.06	Diagnosed Febrile Response

Execution of the learning techniques is highly dependent on the nature of the training data. Confusion matrices are very useful for evaluating classifiers. To evaluate the robustness of classifier, the usual methodology is to perform cross validation on the classifier. The columns represent the predictions, and the rows represent the actual class.

In this simple experiment, from Table I, we can say that a Sacking, J48 requires around 0.05 seconds compared to Naive Bayes which requires around 0.02 seconds. Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to

distinguish between the reliability of the data collected and their validity. The average Kappa score from the selected algorithm is around 0.08-0.40. Based on the Kappa Statistic criteria, the accuracy of this classification purposes is substantial. This experiment implies a very commonly used indicator which is mean of absolute errors and root mean squared errors. Alternatively, the relative errors are also used. Since, we have two readings on the errors, taking the average value will be wise.

VI. CONCLUSION

In medical diagnosis various data mining techniques are available. In this study, for classification of medical data we employed Sacking algorithm because it produce human readable classification rules which are easy to interpret. Researchers have been investigating applying different data mining techniques to help health care professionals in the diagnosis of febrile response. Sacking algorithm is one of the successful data mining techniques used in the diagnosis of febrile response patients. This paper investigates experiments are conducted to find the best classifier for predicting the diagnosis of type of febrile response patients. This paper systematically investigates applying different methods of classifier technique in the diagnosis of heart disease patients. The results show that Sacking algorithm accuracy of 85.03% and the total time taken to build the model is at 0.05 seconds in the diagnosis of pyrexia patients. Finally, some limitations on this work are noted as pointers for future research.

The empirical results show that we can produce short but accurate prediction list for the pyrexia patients by applying the predictive models to the records of incoming new patients. This study will also work to identify those patients which needed special attention.

REFERENCES

[1] Global Burden of Disease. 2004 update (2008). “World Health Organization”.

[2]Coronary Heart Diseases in India. Mark D Huffman. “Center for Chronic Disease Control”.http://sancd.org/uploads/pdf/factsheet_CHD.pdf

[3] L. Breiman, “Sacking predictors”, Machine Learning, 26, 1996, 123-140.

[4] Cheung, N., “Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering, B.Sc. Thesis, University of Queenland.”, 2001.

[5] Tsirogiannis, G.L, Frossyniotis, D, Stoitsis, J, Golemati, S, Stafylopatis, A Nikita,K.S,”Classification of Medical Data with a Robust Multi-Level Combination scheme”, IEEE international joint Conference on Neural Networks.

[6] Kaewchinporn .C, Vongsuchoto. N, Srisawat. A ” A Combination of Decision Tree Learning and Clustering for Data Classification”, 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE).

[7] Liu Ya-Qin, Wang Cheng, Zhang Lu,” Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data” , 3rd International Conference on Bioinformatics and Biomedical Engineering , 2009.

[8] Sitar-Taut, V.A., et al., “Using machine learning algorithms in pyrexiadisease risk evaluation”. Journal of Applied Computer Science & Mathematics, 2009.

[9] Wu, X., et al., “Top 10 algorithms in data mining analysis. Knowl. Inf. Syst.”, 2007.

[10] S. K. Yadev & Pal., S. 2012. Data Mining: “A Prediction for Performance Improvement of Engineering Students using Classification, World of Computer Science and Information Technology” (WCSIT), 2(2), 51-56.

[11] Kappa at http://www.dmi.columbia.edu/homepages/chuangj/ kappa.

BIOGRAPHY

V.Sathya Preiya received her B.Sc degree in Computer Science from C.P.A College, Madurai Kamaraj University, Madurai, India in 1998 and M.C.A degree in Computer Science and Applications from V.V.Vanniaperumal College for Women, Madurai Kamaraj University, Madurai, India in 2001 and M.Phil degree in Computer Science, Bharathidasan University, Trichirapalli, India in 2007 and M.E Computer Science and Engineering in Sathyabama University, Chennai, India in 2010. Currently she is working as Associate Professor in Panimalar Engineering College, Chennai, India. She has 14 years of teaching experience. Her interested area in the field of Data Structures, Object Oriented Programming, Algorithmics and Data Analytics and Big Data.

M.Sangeetha received her M.Sc degree in Information Technology from P.R College, Bharathidasan University, Trichy, India in 2002 and M.Tech degree in Computer Science and Engineering from M.G.R Engineering College, Dr.M.G.R. University, Chennai, India in 2007. Currently she is working as Assistant Professor in Panimalar Engineering College, Chennai, India. She has 10 years of teaching experience. Her interested area in the field of Computer networks, software engineering, Data structures and cloud computing to the distributed parallel architecture based Systems.

S.T.Santhana lakshmi received her B.E degree in Computer Science and Engineering from V.L.B Janakiammal College of engineering & technology, Bharathiyar University, Chennai, India in 2003 and M.Tech degree in Information Technology from Sathyabama University, Chennai, India in 2009. Currently she is working as Associate Professor in Panimalar Engineering College, Chennai, India. She has 11 years of teaching experience. Her interested area in the field of Artificial Intelligence, Data structures, Compiler design, Networking and Operating systems.

DOI: 10.15680/IJIRCCE.2016. 0405165